This note gives more of the details regarding the double bootstrap approach used to estimate the distribution of

$$\tau_{r_G} = \frac{\sqrt{n}\left(\bar{d}_{r_G} - \mu_{r_G}\right)}{\sqrt{2}\, s_{r_G}}. \tag{1}$$

Because μ_{r_G} is unknown a simple bootstrap approach would be to treat r^*_Glike r_G, $\bar{d}_{r^*_G}$ like μ_{r_G}, $s^*_{r^*_G}$ like s_{r_G}, and $\bar{d}^*_{r^*_G}$ like $\bar{d}_{r_G}$ and estimate

$$\tau^*_{r^*_G} = \frac{\sqrt{n}\left(\bar{d}^*_{r^*_G} - \bar{d}_{r^*_G}\right)}{\sqrt{2}\, s^*_{r^*_G}} \tag{2}$$

where the * terms indicate these are observed values from a bootstrap generated sample. As all terms in (2) are known one can generate many bootstrap samples and empirically estimate the distribution of $\tau^*_{r^*_G}$.

In practice this basic approach tends to underestimate the degree of overestimation bias and some modifications using a double bootstrap approach can improve the situation somewhat. A *bias correction* and a *nested percentile* approach using this double bootstrap were both examined – the results in the paper correspond to the nested percentile approach.

Let $X^*_{ij}, j = 1 \ldots G$ denote the values on the i^{th} array chosen in a bootstrap resampling procedure for one group and $Y^*_{ij}, j = 1 \ldots G$ denote the analogous resampling in the second group. Given a bootstrap sample one can obtain a second nested bootstrapped sample with typical elements X^{**}_{ij} and Y^{**}_{ij}. The terms $\bar{d}^{**}_j$, and s^{**}_j are group differences and pooled standard deviations calculated from the nested bootstrap sample. Analogously, r^{**}_G is the index of the largest of the $t-$tests

$(t_1^{**}, t_2^{**}, \ldots, t_G^{**})$ that is computed and

$$\tau^{**}_{r^{**}_G} = \frac{\sqrt{n}\left(\bar{d}^{**}_{r^{**}_G} - \bar{d}^{*}_{r^{**}_G}\right)}{\sqrt{2}\, s^{**}_{r^{**}_G}}.$$

A simple bootstrap bias correction can be implemented by approximating the bias between F^* and F by the estimated bias between F^{**} and F^* where

$$F \text{ is distribution of } \tau_{r_G} = \frac{\sqrt{n}\left(\bar{d}_{r_G} - \mu_{r_G}\right)}{\sqrt{2}\, s_{r_G}}, \tag{3}$$

$$F^* \text{ is distribution of } \tau^{*}_{r^{*}_G} = \frac{\sqrt{n}\left(\bar{d}^{*}_{r^{*}_G} - \bar{d}_{r^{*}_G}\right)}{\sqrt{2}\, s^{*}_{r^{*}_G}}, \text{ and} \tag{4}$$

$$F^{**} \text{ is distribution of } \tau^{**}_{r^{**}_G} = \frac{\sqrt{n}\left(\bar{d}^{**}_{r^{**}_G} - \bar{d}^{*}_{r^{**}_G}\right)}{\sqrt{2}\, s^{**}_{r^{**}_G}}. \tag{5}$$

Given a bootstrap sample, say the l^{th} of L bootstrapped samples, and associated value $\tau^{*}_{r^{*}_G,l}$ one can compute M nested bootstrap samples with associated values $\tau^{**}_{r^{**}_G,l_m}$ for $m = 1 \ldots M$. Then an estimate of the bias between F^* and F^{**} can be obtained by computing

$$B = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{1}{M}\sum_{m=1}^{M}\tau^{**}_{r^{**}_G,l_m} - \tau^{*}_{r^{*}_G,l}\right). \tag{6}$$

Given that the basic bootstrap estimate F^* is derived from the empirical distribution of

$$\{\tau^{*}_{r^{*}_G,1}, \tau^{*}_{r^{*}_G,2}, \ldots, \tau^{*}_{r^{*}_G,L}\} \tag{7}$$

a bias corrected version F^*_B may be given by the empirical distribution of

$$\{\tau^{*}_{r^{*}_G,1} - B, \tau^{*}_{r^{*}_G,2} - B, \ldots, \tau^{*}_{r^{*}_G,L} - B\} \tag{8}$$

A second type of correction is similar in spirit but seeks to correct for the percentiles associated with confidence intervals rather than just the overall average. The discussion here follows that presented in Chapter 5 of Davison and Hinkley, 1997 and is only sketched here. For a given $\alpha \in (0, 1)$ we are ideally interested in determining F_α^{-1} that satisfies

$$Pr\left(\tau_{r_G} \leq F_\alpha^{-1}\right) = \alpha \tag{9}$$

however we can only observe the bootstrap approximation F_α^{*-1}. Because there may exist some bias in the bootstrap version it is likely

$$Pr\left(\tau_{r_G} \leq F_\alpha^{*-1}\right) \neq \alpha \tag{10}$$

and we would like to determine a corrected percentile, $q(\alpha)$ that satisfies

$$Pr\left(\tau_{r_G} \leq F_{q(\alpha)}^{*-1}\right) = \alpha. \tag{11}$$

As before, estimating this correction requires a nested bootstrap procedure. The estimate of the correction, $\hat{q}(\alpha)$ satisfies

$$Pr^*\left(\tau_{r_G^*}^* \leq F_{\hat{q}(\alpha)}^{**-1}\right) = \alpha \tag{12}$$

where Pr^* indicates the probability is taken with respect to the bootstrap distribution obtained by resampling and F^{**-1} is the empirical distribution in (5). To obtain $\hat{q}(\alpha)$ we again suppose there are L initial bootstrap samples and M nested bootstrap samples for each of the L initial samples. Define

$$u_l^* = \frac{1}{M} \sum_{m=1}^{M} I\{\tau_{r_G^*,l}^* \leq \tau_{r_G^{**},l_m}^{**}\} \tag{13}$$

(where $I\{\}$ is an indicator function taking the value 1 when $\tau^{*}_{r^{*}_{G},l} \leq \tau^{**}_{r^{**}_{G},l_m}$ and 0 otherwise) and order the obtained values $\{u^*_1, u^*_2, \ldots, u^*_L\}$. Then $\hat{q}(\alpha)$ is the $\alpha \cdot 100th$ percentile of the $\{u^*_l, l = 1 \ldots L\}$, or in other words, the $(L+1)\alpha th$ ordered value of $\{u^*_1, u^*_2, \ldots, u^*_L\}$. Once $\hat{q}(\alpha)$ has been obtained for relevant values of α one can derive an alternative form of a 95% confidence interval, e.g.

$$\mu_{r_G} \in \left[\bar{d}_{r_G} - s_{r_G} \frac{\sqrt{2}}{\sqrt{n}} F^{*-1}_{\hat{q}(.975)},\ \bar{d}_{r_G} - s_{r_G} \frac{\sqrt{2}}{\sqrt{n}} F^{*-1}_{\hat{q}(.025)} \right]. \tag{14}$$

This nested percentile method and the bias correction method indicated above in (8) are compared to the basic bootstrap method (as described in (4) i.e. without correction from a second order level of bootstrapping) and traditional $t-$statistic approach in Table 1. The table is a more comprehensive examination than Table 1 presented in the paper. Results correspond to 1000 simulations; each simulation corresponds to a realization of X_{ij} and Y_{ij} values and is associated with $L = 500$ first level bootstrap simulations and $M = 250$ second level bootstrap values.

Table 1 indicates that all the bootstrap methods have about the same coverage probabilities for the 90% and 95% interval regions though the basic bootstrap appears to fare worse for the 50% and 80% intervals. All perform much better than the naïve approach. A second section of the table shows that though overall coverage probabilities may be good for the two methods using a second level bootstrap, these intervals still fail to do a good job at eliminating all the overestimation bias. For instance, for both the bias corrected and nested percentile methods the true value of μ_{r_G} lies in the region $\left(-\infty, \bar{d}_{r_G} - s_{r_G} \frac{\sqrt{2}}{\sqrt{n}} F^{*-1}_{.75}\right)$ 39% of the time instead of the desired rate of 25%. This indicates that point estimates, perhaps based upon $\bar{d}_{r_G} - s_{r_G}\frac{\sqrt{2}}{\sqrt{n}}F^{*-1}_{.50}$, are likely to still overestimate μ_{r_G}. In the simulations with 444 variables this asymmetry is more severe for the basic bootstrap leading to the recommendation to use either of the methods requiring a second level of bootstrapping.

Table 2 provides similar information for the methods when only 10 variables are used and the degree of overestimation is somewhat less. Here the nested percentile method has the best coverage though the length of the intervals are 10-15% longer than those for the other bootstrap methods. Also, the bias corrected version does not fare as well in terms of coverage probabilities though it and the nested percentile appear to exhibit less asymmetry in the extreme regions. From these results and those using 444 variables it appears the nested percentile method may be preferable though it typically has the widest intervals.

Table 3 completes the more comprehensive examination with the case when overestimation is not likely to occur. Here we see that all methods perform approximately equally well though the nested method shows perhaps some degree of asymmetry and has confidence intervals that are slightly wider. However, given its relatively good performance over the other two simulation situations it may still be the most appropriate choice among those presented here.

References

Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and their Application.* London: Cambridge University Press.

Coverage Characteristics

Nominal Interval	Basic Coverage and Mean Length	Bias Corrected Coverage and Mean Length	Nested Percentile Coverage and Mean Length	Naïve $t-$statistic Coverage and Mean Length
$25^{th} - 75^{th}$	41%, .46	50%, .46	50%, .48	3%, .39
$10^{th} - 90^{th}$	77%, .90	81%, .90	81%, .90	12%, .75
$5^{th} - 95^{th}$	90%, 1.18	92%, 1.18	91%, 1.18	22%, .97
$2.5^{th} - 97.5^{th}$	96%, 1.44	96%, 1.44	96%, 1.42	36%, 1.17

Asymmetry Characteristics

Nominal Intervals	Basic Coverage	Bias Corrected Coverage	Nested Percentile Coverage	Naïve $t-$statistic Coverage
$0^{th} - 25^{th}, 75^{th} - 100^{th}$	53%, 6%	39%, 11%	39%, 11%	96%, 1%
$0^{th} - 10^{th}, 90^{th} - 100^{th}$	22%, 1%	15%, 4%	15%, 4%	88%, 0%
$0^{th} - 5^{th}, 95^{th} - 100^{th}$	9%, 1%	7%, 1%	8%, 1%	78%, 0%
$0^{th} - 2.5^{th}, 97.5^{th} - 100^{th}$	4%, 0%	3%, 1%	3%, 1%	64%, 0%

Table 1. Confidence Interval Characteristics for μ_{r_G} with n=14, G=444, Effect Sizes Evenly spaced in $(0, 2]$, Variables Independent, 1000 simulations

Coverage Characteristics

Nominal Interval	Basic Coverage and Mean Length	Bias Corrected Coverage and Mean Length	Nested Percentile Coverage and Mean Length	Naïve $t-$statistic Coverage and Mean Length
$25^{th} - 75^{th}$	45%, .44	43%, .44	49%, .49	34%, .48
$10^{th} - 90^{th}$	76%, .85	71%, .85	77%, .96	54%, .93
$5^{th} - 95^{th}$	87%, 1.10	83%, 1.10	88%, 1.25	74%, 1.21
$2.5^{th} - 97.5^{th}$	92%, 1.33	90%, 1.33	93%, 1.53	84%, 1.45

Asymmetry Characteristics

Nominal Intervals	Basic Coverage	Bias Corrected Coverage	Nested Percentile Coverage	Naïve $t-$statistic Coverage
$0^{th} - 25^{th}, 75^{th} - 100^{th}$	37%, 18%	31%, 26%	28%, 23%	75%, 1%
$0^{th} - 10^{th}, 90^{th} - 100^{th}$	16%, 8%	15%, 14%	13%, 10%	46%, 0%
$0^{th} - 5^{th}, 95^{th} - 100^{th}$	9%, 4%	8%, 9%	7%, 6%	26%, 0%
$0^{th} - 2.5^{th}, 97.5^{th} - 100^{th}$	5%, 3%	5%, 5%	4%, 3%	16%, 0%

TABLE 2. Confidence Interval Characteristics for μ_{r_G} with n=14, G=10, Effect Sizes Evenly spaced in $(0, 1]$, Variables Independent, 1000 simulations

Coverage Characteristics

Nominal Interval	Basic Coverage and Mean Length	Bias Corrected Coverage and Mean Length	Nested Percentile Coverage and Mean Length	Naïve $t-$statistic Coverage and Mean Length
$25^{th} - 75^{th}$	48%, .51	48%, .51	48%, .52	50%, .51
$10^{th} - 90^{th}$	77%, .98	78%, .98	79%, 1.00	78%, 1.0
$5^{th} - 95^{th}$	88%, 1.28	89%, 1.28	89%, 1.32	89%, 1.27
$2.5^{th} - 97.5^{th}$	93%, 1.55	94%, 1.55	94%, 1.61	94%, 1.54

Asymmetry Characteristics

Nominal Intervals	Basic Coverage	Bias Corrected Coverage	Nested Percentile Coverage	Naïve $t-$statistic Coverage
$0^{th} - 25^{th}, 75^{th} - 100^{th}$	26%, 26%	26%, 26%	27%, 25%	25%, 25%
$0^{th} - 10^{th}, 90^{th} - 100^{th}$	12%, 10%	12%, 10%	13%, 8%	11%, 11%
$0^{th} - 5^{th}, 95^{th} - 100^{th}$	6%, 6%	6%, 6%	7%, 4%	5%, 6%
$0^{th} - 2.5^{th}, 97.5^{th} - 100^{th}$	3%, 4%	3%, 3%	4%, 2%	3%, 3%

TABLE 3. Confidence Interval Characteristics for μ_{r_G} with n=14, G=10, Effect Sizes = {3,0,0,0,0,0,0,0,0,0}, Variables Independent, 1000 simulations